RR Project1

Shengbai Zhang

7/13/2021

Load and Preprocessing

```
activity <- read.csv("~/Downloads/activity.csv")
head(activity)
```

```
##   steps       date interval
## 1    NA 2012-10-01        0
## 2    NA 2012-10-01        5
## 3    NA 2012-10-01       10
## 4    NA 2012-10-01       15
## 5    NA 2012-10-01       20
## 6    NA 2012-10-01       25
```

```
dim(activity)
```

```
## [1] 17568     3
```

```
summary(activity)
```

```
##      steps            date              interval     
##  Min.   :  0.00   Length:17568       Min.   :   0.0  
##  1st Qu.:  0.00   Class :character   1st Qu.: 588.8  
##  Median :  0.00   Mode  :character   Median :1177.5  
##  Mean   : 37.38                      Mean   :1177.5  
##  3rd Qu.: 12.00                      3rd Qu.:1766.2  
##  Max.   :806.00                      Max.   :2355.0  
##  NA's   :2304
```

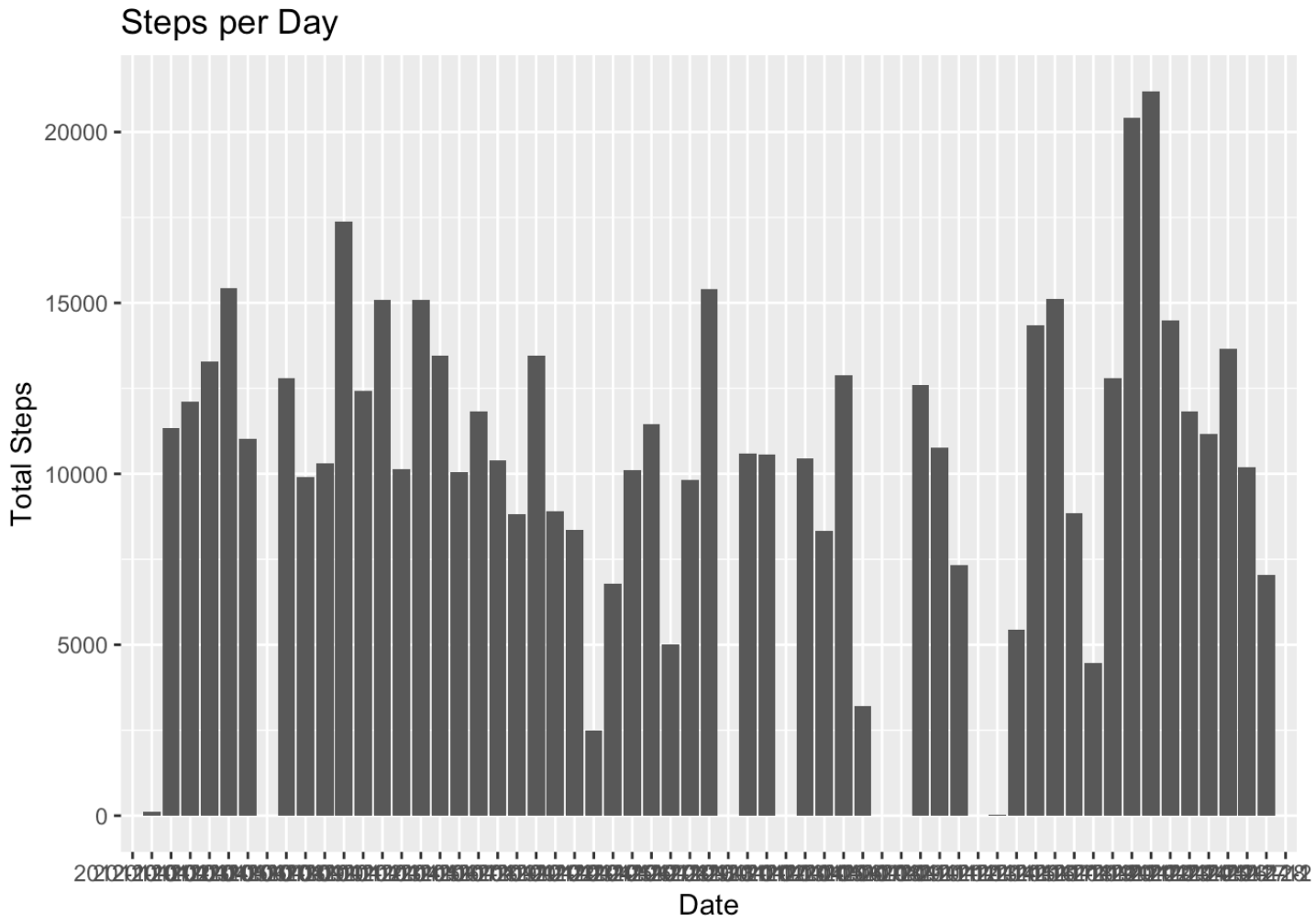
```
activity$date<-as.Date(activity$date,"%Y-%m-%d")
```

Mean Total Steps Taken Per Day

```
a1<- data.frame(tapply(activity$steps,activity$date,sum,na.rm = T))
a1$date<-rownames(a1)
rownames(a1)<-NULL
library(ggplot2)
plot1<-ggplot(a1,aes(x=a1$date,y=a1$tapply.activity.steps..activity.date..sum..na.rm.
..T.))+geom_bar(stat="identity")+ylab("Total Steps")+xlab("Date")+ggtitle("Steps per
Day")
plot1
```

```
## Warning: Use of `a1$date` is discouraged. Use `date` instead.
```

```
## Warning: Use of `a1$tapply.activity.steps..activity.date..sum..na.rm...T.`
## is discouraged. Use `tapply.activity.steps..activity.date..sum..na.rm...T.`
## instead.
```




```
mean(a1[,1])
```

```
## [1] 9354.23
```

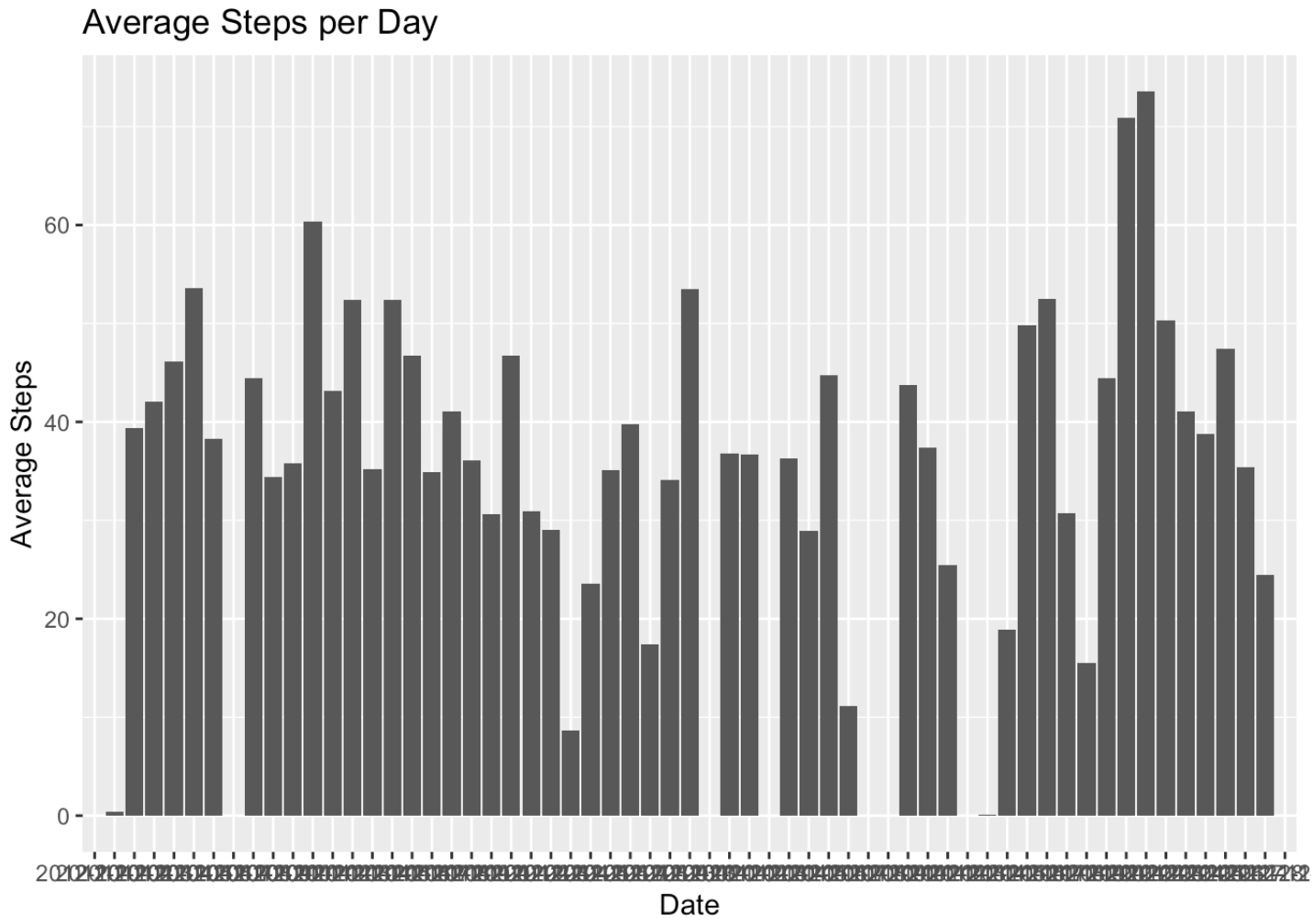
```
median(a1[,1])
```

```
## [1] 10395
```

Average Daily Pattern

```
a2<-data.frame(tapply(activity$steps,activity$date,mean,na.rm=T))
a2$date<-rownames(a2)
rownames(a2)<-NULL
a2$average<-a2$tapply.activity.steps..activity.date..mean..na.rm...T.
ggplot(a2,aes(x=date,y=average))+ylab("Average Steps")+xlab("Date")+ggtitle("Average Steps per Day")+ geom_bar(stat="identity")
```

```
## Warning: Removed 8 rows containing missing values (position_stack).
```




```
activity[which.max(activity$steps),]
```

```
##       steps       date interval
## 16492   806 2012-11-27      615
```

2012-11-27 6:15 contain max steps

Imputing Missing Data

The presence of missing days may introduce bias into some calculations or summaries of the data.

And the way to solve this problem is to impute values. Common imputations used include constant, regression model output, or mean value.

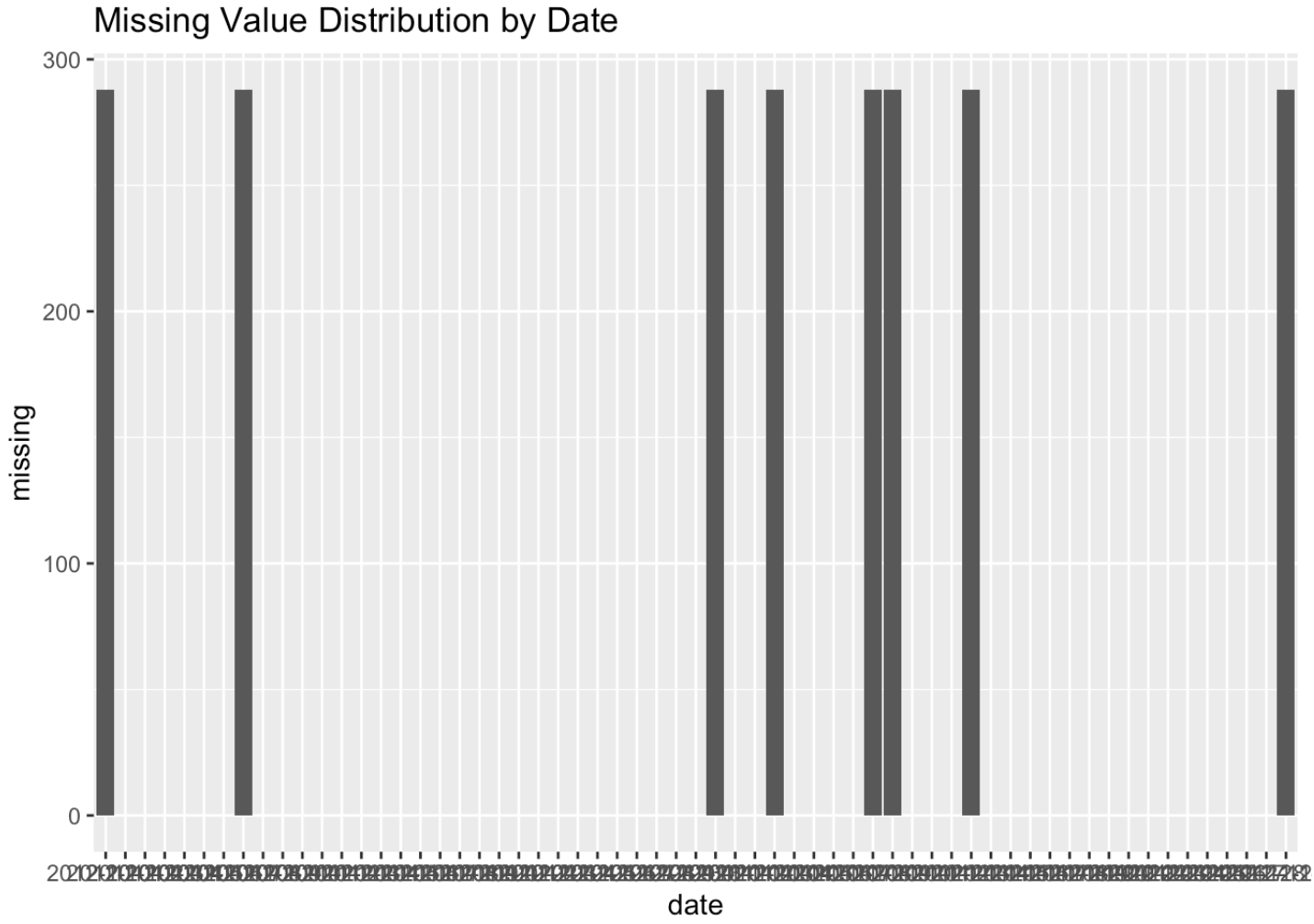
For simplicity, mean imputation will be used.

```
a3<-activity
a3$missing<-is.na(a3$steps)
a3<-aggregate(data=a3, missing~date+interval,FUN = "sum")
a31<-data.frame(tapply(a3$missing,a3$date,sum))
a31$date<-rownames(a31)
rownames(a31)=NULL
names(a31)<-c("missing","date")
```

```
a32<-data.frame(tapply(a3$missing,a3$interval,sum))
a32$date<-rownames(a32)
rownames(a32)<-NULL
names(a32)<-c("missing","Interval")
```

plots

```
ggplot(a31,aes(x=date,y=missing))+ggtitle("Missing Value Distribution by Date")+
  geom_bar(stat = "identity")
```



From the plot, we could observe that there are 8 days that have no steps value, we donot know what

happened at those days, but there is a pattern, and the mean imputation is desirable